Published by Oxford University Press *Nucleic Acids Research, 2002, Vol. 30, No. 1* ***17–20***

GenBank

Dennis A. Benson*, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell, Barbara A. Rapp and David L. Wheeler

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Building 38A, 8600 Rockville Pike, Bethesda, MD 20894, USA

Received September 20, 2001; Accepted October 10, 2001

ABSTRACT

The GenBank sequence database incorporates publicly available DNA sequences of more than 105 000 different organisms, primarily through direct submission of sequence data from individual laboratories and large-scale sequencing projects. Most submissions are made using the BankIt (web) or Sequin programs and accession numbers are assigned by GenBank staff upon receipt. Data exchange with the EMBL Data Library and the DNA Data Bank of Japan helps ensure comprehensive worldwide coverage. GenBank data is accessible through NCBI's integrated retrieval system, Entrez, which integrates data from the major DNA and protein sequence databases along with taxonomy, genome, mapping, protein structure and domain information, and the biomedical literature via PubMed. Sequence similarity searching is provided by the BLAST family of programs. Complete bimonthly releases and daily updates of the GenBank database are available by FTP. NCBI also offers a wide range of World Wide Web retrieval and analysis services based on GenBank data. The GenBank database and related resources are freely accessible via the NCBI home page at http://www.ncbi.nlm.nih.gov.

INTRODUCTION

GenBank (1) is a public database of all known nucleotide and protein sequences with supporting bibliographic and biological annotation, built and distributed by the National Center for Biotechnology Information (NCBI), a division of the National Library of Medicine (NLM), located on the campus of the US National Institutes of Health (NIH) in Bethesda, MD.

NCBI builds GenBank primarily from the direct submission of sequence data from authors. Another major source of data is bulk submission of EST, GSS and other high-throughput data from sequencing centers. The US Office of Patents and Trademarks (USPTO) also contributes sequence data from issued patents. The data are supplemented by sequences submitted to other public databases. Through a long-standing international collaboration with the EMBL Data Library (2) in the UK and the DNA Databank of Japan (DDBJ) (3), data are exchanged daily to ensure that all three sites maintain a comprehensive collection of sequence information. NCBI makes the data available at no cost over the Internet, by FTP access and by web text and sequence similarity search services.

NCBI also offers a wide range of World Wide Web retrieval and analysis services which operate on the GenBank data (4).

ORGANIZATION OF THE DATABASE

GenBank continues to grow at an exponential rate. Over the past 12 months 4.6 million new sequences have been added. As of Release 125 in August 2001, GenBank contained over 13.5 billion nucleotide bases from 12.8 million different sequences. Complete genomes (http://www.ncbi.nlm.nih.gov/Genomes/index.html) represent a growing portion of the database, with 35 of the 57 complete microbial genomes now in GenBank deposited over the past 2 years. Recent additions include *Streptococcus pneumoniae* strain R6 and *Agrobacterium tumefaciens*, a common soil bacterium that causes crown gall disease by integrating some of its DNA into its plant host. The nearly complete eukaryote genomes of *Drosophila melanogaster*, *Caenorhabditis elegans*, *Arabidopsis thaliana* and *Homo sapiens* now join the list. There are at least 57 additional microorganism genomes, plus those of seven eukaryotes including *Plasmodium falciparum* 3D7, that are being sequenced. Many of these are expected to be in the public databases over the coming year. Historically, GenBank had been doubling in size about every 18 months, but that rate has accelerated to doubling every 15 months due primarily to the enormous growth in data from expressed sequence tags (ESTs). Over 67% of the sequences in GenBank Release 125 are ESTs, and current EST projects for human, mouse, rat and other organisms will contribute still more data.

Sequence-based taxonomy

Over 105 000 different species are represented in GenBank and new species are being added at the rate of over 1400 per month. Human sequences constitute 38% of the total sequences (29% of all sequences are human ESTs). After *H.sapiens*, the top species in GenBank in terms of the number of bases include *Mus musculus*, *D.melanogaster*, *C.elegans* and *A.thaliana*. Database sequences are processed and can be queried using a comprehensive sequence-based taxonomy (http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html) developed by NCBI in collaboration with EMBL and DDBJ

*To whom correspondence should be addressed. Tel: +1 301 435 5980; Fax: +1 301 480 9241; Email: dab@ncbi.nlm.nih.gov

and with the valuable assistance of external advisors and curators.

GenBank records and divisions

Each GenBank entry includes a concise description of the sequence, the scientific name and taxonomy of the source organism, bibliographic references and a table of features (http://www.ncbi.nlm.nih.gov/collab/FT/index.html) that identifies coding regions and other sites of biological significance, such as transcription units, repeat regions, sites of mutations or modifications, and other sequence features. Protein translations for coding regions are also in the feature table.

The files in the GenBank distribution have traditionally been divided into 'divisions' that roughly correspond to taxonomic divisions, e.g. bacteria, viruses, primates and rodents. In recent years divisions have been added as needed for specific initiatives in biology, such as divisions for EST sequences, genome survey sequences, and high throughput genomic (HTG) sequences. There are currently 17 divisions; a new division, the high-throughput cDNA (HTC) division, for unfinished high-throughput cDNA, has recently been added. For convenience in file transfer, the larger divisions, e.g. EST and primate, are divided into multiple files when posting the bimonthly GenBank releases on NCBI's FTP site.

EST data

ESTs continue to be the major source of new sequence records and genes. Last year there were 5 462 530 sequences in the EST division of GenBank. Over the past year the number of ESTs has increased by >58% to the current total of 8 643 630 sequences representing over 350 different organisms. The top five organisms include: human, with 3 760 298 sequences (44% of the total); mouse, with 2 098 292 sequences (24%); rat, with 316 356 sequences (4%); fruit fly, with 187 283 sequences (2%); and soybean, with 185 925 sequences (2%).

ESTs also continue to provide the major source of new gene discoveries. As part of its daily processing of EST data, NCBI identifies through BLAST searches all homologies for new EST sequences and incorporates that information into the companion dbEST database (5). In order to organize the EST data in a useful fashion, NCBI maintains the UniGene (http://www.ncbi.nlm.nih.gov/UniGene/) collection of unique human (6), mouse, rat, cow, frog, zebrafish, rice, wheat, barley and corn genes. Additional information about UniGene is included in a separate article in this issue (4).

Sequence-tagged site (STS) data

The STS division of GenBank currently contains over 111 635 sequences and includes anonymous STSs based on genomic sequence as well gene-based STSs derived from the 3′ ends of genes and ESTs. These STS records usually include primer sequences, annotations and PCR reaction conditions.

The ultimate purpose for creating high resolution physical maps of the human genome is to create a scaffold for organizing large scale sequencing (7). Physical maps based on STS landmarks are used to develop so-called 'sequence-ready' clones consisting of overlapping cosmids or BACs. As the HTG sequence data derived from these clones are submitted to GenBank, STSs become crucial reference points for organizing, presenting and searching the data. NCBI uses 'electronic PCR' to compare all human sequences with the contents of the STS division of GenBank; this identifies primer-binding sites on the human sequences that may be amplified in a PCR reaction. This tool permits the assignment of an initial location on the map for sequence data and the association of existing GenBank entries to the new reference sequence. The electronic PCR tool is available on the web to enable any researcher with a new human sequence to relate that sequence to existing maps and HTG sequence data.

Genome survey sequence (GSS) data

The GSS division of GenBank continues to grow rapidly, having increased by 50% to a total of 2 681 177 records with >1.4 billion nucleotides. GSS records represent 'random' genomic sequences, but are predominantly represented by 'BAC ends' which are single reads from bacterial artificial chromosomes used in a variety of genome sequencing projects, notably that of human (870 098 records), *Oryza sativa* (93 118 records) and *Tetradon nigroviridis* (188 963 records). The human data is being used (www.ncbi.nlm.nih.gov/genome/clone) along with the STS records in tiling the BACs used for the Human Genome Project (8).

HTG data

The HTG sequences in the HTG division of GenBank are unfinished large-scale genomic records that are in transition to a finished state, after which they will be placed in the appropriate organism division (9). These records are designated as Phase 0–3 depending on the quality of the data. Phase 0 records consist of survey sequences generated to characterize clones and may or may not progress to Phase 1. Phase 1 records contain unfinished sequence, and may consist of unordered, unoriented contigs with gaps. Phase 2 records contain unfinished sequence as ordered, oriented contigs, with or without gaps. Phase 3 records consist of finished sequence, with no gaps and may have annotations. When a HTG record reaches Phase 3 it is moved from the HTG division into the appropriate taxonomic division of GenBank. It is now clear that a great number of human sequences will remain in the unfinished (HTG) division of GenBank as working draft sequence, while completed sequences will continue to move to the corresponding taxonomic division (PRI). Together these two divisions have added some 4000 Mb of new genomic sequences from US-sponsored laboratories over the last 2 years.

HTC data

A new GenBank HTC division has been created for high-throughput cDNA sequencing data. HTC sequences may be single-pass sequences that may have 5′-UTRs and 3′-UTRs at their ends, partial coding regions, and introns. HTC sequences which are finished and of high-quality will be moved to the appropriate taxonomic GenBank division. As of August 2001, GenBank contained 21 326 HTC sequences totalling 26 832 389 bases. The bulk of these HTC sequences, 91%, are from *M.musculus* while ~8% are from *H.sapiens*. A recent project generating HTC data has been described (10).

Sequence identifiers and accession numbers

Each GenBank DNA sequence record is assigned an accession number, which is a stable and unique identifier for the GenBank entry as a whole, and does not change, even when there is a change to the sequence or annotation. In order to

identify specific sequences from different sources, as well as keep track of modifications to the actual sequence data, NCBI additionally assigns a unique identifier, termed a 'gi' number, to each sequence. When a change in a sequence occurs, a new gi number is assigned to the new sequence version. These gi numbers appear on the VERSION line of GenBank records following the accession number.

By agreement among the collaborative DNA sequence databases, a third identifier was introduced in February 1999 which consolidates the information present in both the gi and accession numbers. GenBank displays this identifier on the VERSION line, which appears below the ACCESSION line in the GenBank flat file format and is of the form 'Accession.version'. For example, an entry appearing in the database for the first time has a VERSION number equivalent to the ACCESSION number followed by '.1' to reflect that this is the first version of the sequence in this entry, e.g.

ACCESSION AF000001
VERSION AF000001.1 GI: 987654321

If the nucleotide sequence changes, then so will the gi number and the version, but the accession will remain the same.

A similar system for tracking changes in the corresponding protein translations was also introduced in February 1999. Protein sequences now have identification numbers (in the format of three letters followed by five digits, e.g. AAA00001) that do not change, followed by a version number that increases with each subsequent version of the sequence. This identifier appears as a qualifier for a CDS feature in the FEATURES table portion of a GenBank entry, e.g. /protein_id='AAA00001.1' Protein sequence translations also currently receive their own unique gi number, which appears as a second qualifier on the CDS feature: /db_xref='GI:1233445'.

BUILDING THE DATABASE

The data in GenBank, and the collaborating databases EMBL and DDBJ, come from two sources: (i) individual authors who submit data directly to one of the databases, and (ii) bulk submissions from sequencing centers in the form of ESTs, STSs, GSSs, HTCs or large genomic records (usually sequences from cosmids, BACs or YACs). Data are exchanged daily with DDBJ and EMBL so that the daily updates from NCBI servers incorporate the most recently available sequence data from all sources.

Direct submission

Virtually all records enter GenBank as direct electronic submissions, with the majority of authors using the BankIt or Sequin programs. Many journals require authors with sequence data to submit the data to a public database as a condition of publication.

GenBank staff can usually assign an accession number to a sequence submission within 2 working days of receipt, and do so at a rate of several hundred per day. The accession number serves as confirmation that the sequence has been submitted and allows readers of the article to retrieve the relevant data. All direct submissions receive a systematic quality assurance review including checking for vector contamination, verifying proper translation of coding regions, and checking for correct taxonomy and bibliographic citations. A draft of the GenBank record is passed back to the author for review before it enters the database. Authors have the right to request that their sequences be kept confidential until a date in the future to allow time for the sequence to be published. In these cases, authors are instructed to inform GenBank staff of the publication date of the article in which the sequence is cited in order to ensure a timely release of the data. GenBank policy requires that deposited sequence data be made public when the sequence or accession number is published. Although only the submitting scientist is permitted to modify sequence data or annotations, all users are encouraged to report lags in releasing data or possible errors or omissions to GenBank at update@ncbi.nlm.nih.gov.

Several large-scale sequencing projects are producing megabases of human genomic DNA sequence. NCBI works closely with sequencing centers to ensure timely incorporation of these data into GenBank for public release. In parallel, NCBI has developed methods to integrate these sequences with genetic and physical map data and to search the sequences more effectively (e.g. through options in BLAST to mask Alu and other types of repetitive elements). GenBank offers special batch procedures for large-scale sequencing groups to facilitate data submission, including the program 'fa2htgs' and other tools (11).

BankIt

About 35% of individual submissions are received through a web-based data submission tool, BankIt (http://www.ncbi.nlm.nih.gov/BankIt). With BankIt, authors enter sequence information directly into a form, edit as necessary and add biological annotation (e.g. coding regions, mRNA features). Recent revisions of Bankit allow for the entry of more fielded information through the use of listboxes and pull-down menus. Free-form text boxes allow the submitter to further describe the sequence, without having to learn formatting rules or use restricted vocabularies. Bankit validates submissions, flagging many common errors, and checks for vector contamination using a variant of BLAST called Vecscreen, before creating a draft record in GenBank flat file format for the user to review and revise. BankIt is the tool of choice for simple submissions, especially when only one or a small number of records is submitted (9). BankIt can also be used by submitters to update their existing GenBank records.

Sequin

NCBI has developed a stand-alone multi-platform submission program called Sequin (http://www.ncbi.nlm.nih.gov/Sequin/index.html) which can also be linked online to NCBI. Sequin handles simple sequences (e.g. a cDNA), as well as long sequences and segmented entries, for which BankIt and other web-based submission tools are not well-suited. Sequin has convenient editing and complex annotation capabilities and contains a number of built-in validation functions for enhanced quality assurance. It is also designed to facilitate the submission of sequences from phylogenetic, population studies, mutation studies and environmental samples, and can incorporate alignment data. Sequin can be used to edit and update sequence records, as well as to perform sequence analysis. For example, Sequin can now incorporate any analysis tool available on the web that accepts FASTA or ASN.1 (Abstract Syntax Notation 1) formatted data as its input. In addition, Sequin is able to work on large records (e.g. the *Escherichia coli* genome at 5.6 Mb) and read in all of its

annotations via simple tables. Versions for Macintosh, PC and Unix computers are available via anonymous FTP to ftp.ncbi.nih.gov in the 'sequin' directory. Once a submission is completed, users can email it to the address gb-sub@ncbi.nlm.nih.gov. Additional information about Sequin can be found through the NCBI home page.

RETRIEVING GENBANK DATA

The Entrez system

Entrez (http://www.ncbi.nlm.nih.gov/Entrez/) is an integrated database retrieval system that accesses DNA and protein sequence data, genome mapping data, population sets, phylogenetic sets, environmental sample sets, gene expression data, the NCBI taxonomy, protein domain information, protein structures from the Molecular Modeling Database, MMDB (12), and MEDLINE references via PubMed. The DNA and protein sequence data are integrated from a variety of sources and therefore include more sequence data than are available within GenBank alone. Entrez searching is provided on NCBI's web site, via the Query Email server (query@ncbi.nlm.nih.gov), and as a network client that can be downloaded by FTP. Entrez is also discussed elsewhere in this issue (4).

BLAST sequence-similarity searching

The most frequent type of analysis performed using GenBank is the search for sequences similar to a query sequence. NCBI offers the BLAST (http://www.ncbi.nlm.nih.gov/BLAST/) family of programs to locate good alignments between a query sequence and database sequences (13,14). BLAST searching is provided on NCBI's web site, via an email server (blast@ncbi.nlm.nih.gov), and as a set of stand-alone programs distributed by FTP. BLAST is discussed in more detail in a separate article in this issue (4).

Obtaining GenBank by FTP

NCBI uses the ASN.1 data format for internal maintenance of GenBank, but distributes the GenBank releases in the traditional flat-file format. The full GenBank release (issued every 2 months) and the daily updates (which also incorporate sequence data from EMBL and DDBJ) are available by anonymous FTP from NCBI at ftp.ncbi.nih.gov as well as from two mirror sites, at the San Diego Supercomputer Center (ftp://genbank.sdsc.edu/pub) and at the University of Indiana (http://bio-mirror.net/biomirror/genbank). The full release in flat-file format is available as compressed files in the directory, 'genbank'. A cumulative update file is contained in the sub-directory, 'daily', and a non-cumulative set of updates is contained in 'daily-nc'. A set of sequence-only files in FASTA format, corresponding to the GenBank database subsets searched by BLAST and including the non-redundant nucleotide and protein databases, is available in the 'blast/db' directory.

MAILING ADDRESS

GenBank, National Center for Biotechnology Information, Building 38A, Room 8S-803, 8600 Rockville Pike, Bethesda, MD 20894, USA. Tel: +1 301 496 2475; Fax: +1 301 480 9241.

ELECTRONIC ADDRESSES

NCBI home page, http://www.ncbi.nlm.nih.gov/; submission of sequence data to GenBank, gb-sub@ncbi.nlm.nih.gov; revisions to GenBank entries and notification of release of 'confidential' entries, update@ncbi.nlm.nih.gov; general information about NCBI and services; info@ncbi.nlm.nih.gov.

CITING GENBANK

If you use GenBank as a tool in your published research, we ask that this paper be cited.

REFERENCES

1. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J., Ouellette,B.F.F., Rapp,B.A. and Wheeler,D.L. (2000) GenBank. *Nucleic Acids Res.*, **28**, 15–18.
2. Stoesser,G., Baker,W., van den Broek,A., Camon,E., Garcia-Pastor,M., Kanz,C., Kulikova,T., Lombard,V., Lopez,R., Parkinson,H., Redaschi,N., Sterk,P., Stoehr,P. and Tuli,M.A. (2001) The EMBL nucleotide sequence database. *Nucleic Acids Res.*, **29**, 17–21. Updated article in this issue: *Nucleic Acids Res.* (2002), **30**, 21–26.
3. Tateno,Y., Miyazaki,S., Ota,M., Sugawara,H. and Gojobori,T. (2000) DNA data bank of Japan (DDBJ) in collaboration with mass sequencing teams. *Nucleic Acids Res.*, **28**, 24–26. Updated article in this issue: *Nucleic Acids Res.* (2002), **30**, 27–30.
4. Wheeler,D.L., Church,D.M., Lash,A.E., Leipe,D.D., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M.,Tatusova,T.A., Wagner,L. and Rapp,B.A. (2002) Database resources of the National Center for Biotechnology Information: 2002 update. *Nucleic Acids Res.*, **30**, 13–16.
5. Boguski,M.S., Lowe,T.M. and Tolstoshev,C.M. (1993) dbEST—database for 'expressed sequence tags'. *Nature Genet.*, **4**, 332–333.
6. Schuler,G.D. (1997) Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J. Mol. Med.*, **75**, 694–698.
7. Hudson,T.J., Stein,L.D., Gerety,S., Ma,J., Castle,A.B., Silva,J., Slonim,D.K., Baptista,R., Kruglyak,L., Xu,S.-H. *et al.* (1995) An STS-based map of the human genome. *Science*, **270**, 1945–1954.
8. Smith,M.W., Holmsen,A.L., Wei,Y.H., Peterson,M. and Evans,G.A. (1994) Genomic sequence sampling: a strategy for high resolution sequence-based physical mapping of complex genomes. *Nature Genet.*, **7**, 40–47.
9. Kans,J.A. and Ouellette,B.F.F. (2001) In Baxevanis,A. and Ouellette,B.F.F. (eds), *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins.* John Wiley and Sons, New York, NY, pp. 65–81.
10. Hayashizaki,Y. (2001) Functional annotation of a full-length mouse cDNA collection *Nature*, **409**, 685–690.
11. Ouellette,B.F.F. and Boguski,M.S. (1997) Database divisions and homology search files: a guide for the perplexed. *Genome Res.*, **7**, 952–957.
12. Wang,Y., Geer,L., Madej,T., Marchler-Bauer,A., Zimmerman,D. and Bryant,S.H. (2002) MMDB: 3D structure data in Entrez. *Nucleic Acids Res.*, **30**, 249–252.
13. Altschul,S.F, Madden,T.L., Schaffer,A.A., Zhang,J., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
14. Zhang,Z., Schaffer,A.A., Miller,W., Madden,T.L., Lipman,D.J., Koonin,E.V. and Altschul,S.F. (1998) Protein sequence similarity searches using patterns as seeds. *Nucleic Acids Res.*, **26**, 3986–3991.